Clustering Marketing Datasets with Data Mining Techniques

Özgür Örnek
International Burch University, Sarajevo
oornek@ibu.edu.ba

Abdülhamit Subaşı
International Burch University, Sarajevo
asubasi@ibu.edu.ba

Abstract: Customer analysis is crucial phase for companies in order to create new campaign for their existing customers. If a company can understand customer features and make efforts to fulfill their wants and provide friendly service then the customer will be more supportive to the enterprise. The aim of this study was to develop a methodology to identify the characteristics of customers. It involved identification of the demographic characteristics of customers based on the analysis of categorical data using data mining clustering methods. The extracted knowledge can help companies identify valuable customers, and enable companies to make efficient knowledge-driven decisions.

Keywords: Data mining, Clustering, Marketing Segmentation, K-means, E-M Algorithm

1. Introduction

Customer analysis is crucial phase for companies in order to create new campaign for their existing customers. Companies are able to group or cluster certain customers which have similar features. This may assist companies to make better marketing strategies over certain customer groups. Companies recognize that their existing customer database is their most important asset (Athanassopoulos, 2000). It is also important that how to effectively process and use customer data. Thus, this new techniques to assist analyze, comprehend or even visualize the massive amounts of stored data obtained from business and scientific applications (Liao et al, 2004). Data mining is the process of discovering and extracting considerable customer knowledge, such as rules, patterns, associations, clusters, and significant structures from large amounts of data stored in databases (Liao et al., 2008; Coussement et al., 2010).

According to a research conducted by Reinartz et al., it is more beneficial to keep and satisfy existing customers than to constantly attract new customers who are characterized by a high attrition rate (Reinartz et al., 2003). Thus, if a company can understand customer features and make efforts to fulfill their wants and provide friendly service then the customer will be more supportive to the enterprise. For instance, specific measures and motivation may be proposed to the most risky customer groups, i.e. the most disposed to leave the company, they may remain constant (Burez et al., 2007).

The aim of this study was to develop a methodology to identify the characteristics of customers. It involved identification of the demographic characteristics of customers based on the analysis of categorical data using data mining clustering methods. The extracted knowledge can help companies identify valuable customers, and enable companies to make efficient knowledge-driven decisions.

2. Materials and Methods

In last decades, data mining techniques have been employed to forecast customer behavior (Giudici et al., 2002). Data mining is an application that involves specific algorithms for pattern extraction (Mitra et al., 2001). Data mining implements association algorithm according to decision attributes in order to analyze customer features so that the marketing managers can develop strategies for target customers.

2.1. Data mining

Data mining, also known as knowledge discovery in database, is prompted by the need of new techniques to help analyze, understand or even visualize the large amounts of stored data gathered from business and scientific applications. It is the process of investigating knowledge, such as patterns, associations, changes, anomalies or significant structures from large amounts of data stored in database, data warehouse, or other information repositories (Hui et al., 2000). Nowadays, some data mining methods and applications have been developed to analyze the practices and planning methods of sales and marketing management between customers and vendors in the market (Bloemer et al., 2003; Liao et al., 2004)

Another study conducted by Hsieh (Hsieh, 2004) offered a method that integrated data mining and behavioral scoring models for the management of banking customers. He categorized customers into three groups according to their shared behaviors, characteristics, and effectiveness. Marketing managers conclude the profiles of each group of customers and propose management appropriate policies to the characteristics of each group. Customer behavioral variables, demographic variables, and transaction databases are employed to create a method of mining changes in customer behavior in the retail market (Chen et al., 2005). In their study, customer behavior patterns are first recognized using association rule mining. After the association rules for customer behavior are realized, changes in customer behavior are identified by comparing two sets of association rules produced from two datasets from different periods. The changes in patterns can then be investigated and evaluated to provide a basis for creating marketing strategies. Customer behavior analysis in Internet marketing has already been investigated by many researchers (Jenamani et al., 2003). In most of similar researches, data mining technologies are applied to produce a categorized customer profile of the Internet shopper and to further investigate the Web usage pattern of the online consumer. The knowledge obtained through data mining helps to promote informed Internet marketing decision-making and provides for the improvement of Web content and infrastructure to raise Internet marketing (Kwan et al., 2005; Liou et al. 2010).
This paper proposes the clustering analysis for data mining to extract market knowledge of customers' database. In this work we analyzed customer demographic knowledge using clustering techniques, and then relevant knowledge was extracted to explore useful information/knowledge of patterns for marketing and customer relationship management. Knowledge extracted from this analysis can serve as useful input for upper management and analysts of planning and operation and marketing departments.

2.2. Clustering

Clustering is a task of grouping objects into classes of similar objects (Jain et al., 1999). It is an unsupervised classification or partitioning of patterns into groups or clusters based on their locality and connectivity within an n-dimensional space. In this study, clustering has been used for finding clusters of customers with similar characteristics.

2.3. Marketing Data

In this study, we used marketing dataset gathered from shopping mall customers in the San Francisco Bay area (Impact Resources, 1987).

The dataset income data is an extract from this survey. It consists of 14 demographic attributes. This survey's aim was to predict the annual income of household from the other 13 demographics attributes. 8993 instances have been used for this survey. The attributes that are used in this survey summarized as follows:

- Annual income of household (personal income if single)
- Sex
- Marital status
- Age
- Education
- Occupation
- How long have you lived in the san fran./oakland/san jose area?
- Dual incomes (if married)
- Persons in your household
- Persons in household under 18
- Householder status
- Type of home

- Ethnic classification
- What language is spoken most often in your home?

3. Results and Discussion

In this paper, we used Weka software which has some useful advantageous. It is free software system, and it uses the same dataset external representation format. So, it can easily be downloaded from Internet, used without data format problems and, if required, changed using the same programming language (Romero et al., 2007).

Weka (Witten & Frank, 2005) is open source software which contains a collection of machine learning and data mining algorithms for data pre-processing, classification, regression, clustering, association rules, and visualization.

We clustered 3 similar groups from marketing datasets. We used simple K-means and E-M clustering algorithm in Weka system. The K-means algorithm is one of the simplest and most popular clustering algorithms. It is an algorithm that clusters objects based on attributes in k partitions. The Expectation–Maximization (EM) algorithm is developed for incomplete data (Dempster & Laird, 1977). It can be used to run maximum likelihood parameter prediction for mixture models. It applies the principle of maximum likelihood to find the model parameters. The E-M algorithm repeats the Expectation (E) and Maximization (M) steps iteratively after randomly initializing the mixture model parameters. The E and M steps are iterated until an intended convergence is acquired (Witten & Frank, 2000).
We have performed the K-means over the marketing dataset with 3 number of clusters. Weka K-means algorithm results summarized in Table 1 that shows information about the each cluster, the number and percentage of instances in each cluster.

Attribute	**Full Data**	**Cluster0**	**Cluster1**	**Cluster2**
Sex	1.5469	2	1.5974	1
MaritalStatus	3.031	4.2449	1.1208	4.3551
Age	3.4152	2.884	4.2922	2.7799
Education	3.8351	3.4709	4.2199	3.6946
Occupation	3.788	4.2206	3.3105	3.9825
YearsInSf	4.1983	4.207	4.3348	4.003
DualIncome	1.5448	1.0285	2.3122	1.0429
HouseholdMembers	2.8518	2.8443	3.0091	2.6453
Under18	0.6669	0.7052	0.7602	0.499
HouseholdStatus	1.8367	2.1938	1.3345	2.1448
TypeOfHome	1.8557	2.0139	1.5519	2.103
EthnicClass	5.9559	5.843	6.1424	5.8206
Language	1.1275	1.1292	1.103	1.1591
Income	4.895	3.4094	6.6379	4.0859
Clustered Instances	**8993**	2775 (31%)	3587 (40%)	2631 (29%)

Table 1. Weka K-means clustering algorithm results

Second, We have executed the EM over the marketing dataset with number of 3 clusters. Weka E-M algorihtm results have been summarized in Table 2.

Attribute	**Cluster0**	**Cluster1**	**Cluster2**
Sex	1.0544	1.8043	1.5821
MaritalStatus	4.4007	4.2871	1.1166
Age	2.8438	2.8991	4.2066
Education	3.9473	3.4354	4.1412
Occupation	3.5763	4.3277	3.4079
YearsInSf	3.8344	4.2998	4.306

DualIncome	1	1	2.3478
HouseholdMembers	2.341	2.9506	3.0433
Under18	0.2379	0.8005	0.781
HouseholdStatus	2.0581	2.2001	1.3796
TypeOfHome	2.2464	1.9015	1.5975
EthnicClass	5.824	5.8889	6.0904
Language	1.2424	1.0714	1.1157
Income	4.4533	3.5124	6.4129
Clustered Instances	2174 (24%)	3324 (37%)	3495 (39%)

Table 2. Weka EM clustering algorithm results

We can see in Table 1 and Table 2 that there are 3 clusters of customers. According to Table 1, Cluster 0 is characterized by customers with lower or few features. Cluster 1 is characterized by customers with more values than Cluster 0. Finally, Cluster 2 is characterized by customers with values somewhat smaller than cluster 1 but greater than cluster 0. We can also see in the figure that the students are grouped into 3 clusters with regular numbers of customers 2775, 3587 and 2681 respectively.

According to Table 2 results, Cluster 2 has higher values than other clusters, while Cluster 1 has lower values. This information can be used in order to group customers into three types of customers: high valuable customers (cluster 1), lower valuable customers (cluster 2) and non-valuable students (0). Starting from this information, for example, the marketing managers can group customers for making marketing strategies. The marketing managers can also group new customers into these clusters depending on their features.

4. Conclusion

In this study we have conducted data mining clustering techniques over a marketing dataset in order to obtain interesting information in a more efficient and faster way. Marketing managers can use this extracted knowledge to perform relevant strategies over certain customer groups. This paper proposes K-means and E-M algorithm as a methodology of clustering analysis for data mining, which is implemented for mining customer knowledge from the marketing dataset. Knowledge extraction from data mining results is illustrated as knowledge patterns, rules, and knowledge maps in order to propose suggestions and solutions to the case firm for determining marketing strategies.

Three clusters were obtained from the K-means and E-M analysis. Both clustering algorithm results show some characteristic features of customers. These characteristic may briefly explained as follows: customer age range is 35-44, education level is 1 to 3 year college, marital status is married, number of household members is greater than 3, and householder status is own. Briefly, clustering analysis results show that companies can promote a new strategy by considering customers features including age, education, marital status, and dual income. In these regards, the marketing managers can figure out how to maintain its reputation.

References

Athanassopoulos, A. D. (2000) Customer satisfaction *cues to support market segmentation and explain switching behaviour,* Journal of Business Research, 47(3).

Bloemer, J. M. M., Brijs, T., Vanhoof, K., & Swinnwn, G. (2003) *Comparing complete and partial classification for identifying customers at risk*, International Journal of Research in Marketing, 20, 117–131.

Burez, J., & Van den Poel, D. (2007) *CRM at Canal + Belgique: Reducing customer attrition through targeted marketing*, Expert Systems with Applications, 32(2), 277–288.

Chen, M.C., Chiu A.L., Chang H.W. (2005) *Mining changes in customer behavior in retail marketing*, Expert Systems with Applications 28, 773–781.

Coussement, K., Benoit D.F., De Poel, D.V. (2010) *Improved marketing decision making in a customer churn prediction context using generalized additive models*

Dempster, A. K., & Laird, N. M. (1977) *Maximum likelihood from incomplete data via the EM algorithm,* Journal of the Royal Statistical Society, Series B, 39, 1–38.

Giudici, P., Passerone, G. (2002) *Data mining of association structures to model customer behavior*, Computational Statistics and Data Analysis 38, 533-541.

Hsieh, N.C., (2004) *An integrated data mining and behavioral scoring model for analyzing bank customers*, Expert Systems with Applications, 27 623-633.

Hui, S. C., & Jha, G. (2000) *Data mining for customer service support*, Information and Management, 38, 1–13.

Impact Resources, (1987). http://www-stat.stanford.edu/ElemStatLearn.

Jain, A. K., Murty, M. N., Flynn, P. J. (1999) *Data clustering: A review*, ACM Computing Surveys, 31(3), 264–323.

Jenamani, M., Mohapatra, P., Ghose, S. (2003) *A stochastic model of e-customer behavior*, Electronic Commerce Research and Applications 2, 81–94.

Kwan, I., Fong, J., Wong H.K. (2005) An *e-customer behavior model with online analytical mining for internet marketing planning*, Decision Support Systems 41, 189–204.

Liao, S. H., Chen, C. M., & Wu, C. H. (2008) *Mining customer knowledge for product line and brand extension in retailing*, Expert Systems with Applications, 35(3), 1763–1776.

Liao, S. H., Chern, Y. W., & Liao, W. B. (2004) *Information technology and relationship management: a case study of Taiwan's small manufacturing firm*, Technovation, 24, 97–108.

Liao, S. H., Chen, Y. J. (2004) Mining *customer knowledge for electronic catalog marketing*, Expert Systems with Applications 27, 521–532

Liau, J.H.J., Tzeng, G. (2010) A *Dominance-based Rough Set Approach to customer behavior in the airline market*, Information Sciences 180, 2230–2238.

Mitra, S., Mitra, P., Pal, S.K. (2001) *Evolutionary modular design of rough knowledge-based network using fuzzy attributes*, Neurocomputing 36, 45–66.

Reinartz, W. J., & Kumar, V. (2003). *The impact of customer relationship characteristics on profitable lifetime duration*. Journal of Marketing, 67(1).

Romero, C., Ventura, S., Garcia, E. (2007) *Data mining in course management systems: Moodle case study and tutorial*, Computers & Education 51, 368–384.

Witten, I. H., & Frank, E. (2000) *Data mining: Practical machine learning tools and techniques with java implementations*. San Francisco, CA: Morgan Kaufmann.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. Morgan Kaufman.